The approach of the Given Case Study

Dataset was coming from a library named Kitaab. They have shifted from offline mode to digital mode to spread their library facility to people. Since they have to convert everything into digital they have to upload all books online and mark them based on genre. They need help from data science enthusiasts to complete this work faster.

So my first task was to create a model that could detect the book's genre based on the factors given in the dataset. Dataset is quite interesting and very challenging for me to build the model.

Dataset had title, rating, name,num_ratings,num_reviews,num_followers, synopsis, and genre.

So according to the first task, it was a multiclass classification Problem. So first I tried to see unnecessary columns in the dataset. While doing the analysis I got the "**Unnamed: 0**" column which I dropped since it may be the issue with the file or something. Then I tried to see the data types of all features in the dataset. According to the problem statement, the "**num_ratings**" column should be in integer format but It was in object format. The reason behind that was in that column all values were separated using commas and I tried to remove that comma and then Convert that to Integer format the same problem I saw in the "**num_reviews**" column which I solved in the same way. Then I saw in the "**num_followers**" column had the same commas problem but with that, it had "followers" text attached with some of the values of that column which I solved by using str and split function in python and another problem it has "k" in some of the values and I changed it based on the need in the data. I described it clearly in my code notebook.

So after basic cleaning, I focused on the distribution of the data and the correlation in the data. So I got that the **'ratings'** column is almost normally distributed and it had fewer outliers but other numerical columns were very right skewed and had a lot of outliers since the data had almost 1.5k and because of the outliers I drooped rows from 'ratings' column and when I tried to drop others important data loss issue may happen so I kept them. I saw almost a good correlation between "num_reviews" and "num_ratings". I dropped it at first to avoid the multicollinearity issue but after building the model I saw logistics regression was not performing well and I had to go for tree-based algorithms since the multicollinearity issue does not affect the tree or boosting algorithm I kept it.
After that, I dropped all the unnecessary columns - "**title"**, and **"Synopsis".**

Now for the first task target was "**genre**" column so I mapped all genres based on their names a numerical encoding and I did one hot encoding with the **"Name"** column. Then I tried to use GridSearchCv for tuning the parameters of different classification algorithms of Machine Learning. I tried to do the tune manually too and I got Decision Tree Classifier as the best model which I used to do for Classification.

Now in the second task, I had to predict ratings which was a regression problem and I tried to use all possible ML Regression Algorithms to solve the problem but for all the models even after a lot of tuning and possible ways I know to get good accuracy, I was not able to get the best model but still, I took AdaBoost as my final model which was over fitted but still I used that model to complete my task. I am eager to see if I can find any way to solve this problem which I was not able to complete properly.

Thank you.